An amplitude code transmits information at a visual synapse

Article (Accepted Version)

An amplitude code transmitting information at a visual synapse

Ben James[1], Léa Darnet[1], José Moya-Díaz[1], Sofie-Helene Seibel[1] and Leon Lagnado[1]*

School of Life Sciences, University of Sussex, Brighton, UK

[1]These authors contributed equally to this work.

*Corresponding author: l.lagnado@sussex.ac.uk

Most neurons transmit information digitally using spikes that trigger release of synaptic vesicles with low probability. The first stages of vision and hearing are distinct in operating with analogue signals, but it is unclear how these are recoded for synaptic transmission. By imaging the release of glutamate in live zebrafish, we demonstrate that ribbon synapses of retinal bipolar cells encode contrast through changes in both the frequency and amplitude of release events. Higher contrasts caused multiple vesicles to be released within an event, and such coding by amplitude often continued after the rate code had reached a maximum frequency. Glutamate packets equivalent to five vesicles transmitted four times as many bits of information per vesicle compared to those released individually. By discretizing analogue signals into sequences of numbers up to about eleven, ribbon synapses can increase the dynamic range, temporal precision and efficiency with which visual information is transmitted.

The spike code of neurons has been studied in detail[1,2] but much less is known about the vesicle code transmitting information across the synapse[3]. Since the work of Katz[4], the general view has been that changes in presynaptic potential are communicated by modulating the mean rate of a Poisson Process in which vesicles at the active zone are triggered to fuse independently[5,6]. At most synapses, the voltage signal controlling this process is digital, arriving in the form of an all-or-none spike, and the output is also digital, in the form of a vesicle releasing a fixed packet of neurotransmitter. But in the first stages of vision and hearing the neural signal is in analogue form - continuous changes in membrane potential that are graded with the strength of the stimulus[7]. How are these sensory signals recoded for transmission across a synapse?

Synapses driven by graded changes in membrane potential are distinguished from those driven by spikes by the presence of a specialized structure, the ribbon, that holds tens of vesicles just behind the active zone[8,9]. It has been generally assumed that ribbon synapses represent the strength of the incoming stimulus in the form of a rate code generated by changes in the frequency of release events composed of one quantum of neurotransmitter contained within a vesicle[5,10-12]. But it is also known that when these synapses are activated strongly several vesicles can be released within a few milliseconds[13-15]. This process, termed multivesicular release (MVR), is now recognized to be a feature of synaptic connections in several regions of the brain[16]. In ribbon synapses of mechanosensitive hair cells and retinal bipolar cells, fusion of two or more vesicle equivalents can even be synchronized to within 100 μs[17-20], when the process is termed *coordinated* multivesicular release (CMVR) to highlight the deviation from a Poisson process in which vesicles fuse independently of each other. Although the mechanisms underlying MVR are not understood, these observations suggest that the output of a ribbon synapse may not be a binary set of symbols consisting of zero or one vesicle but rather a number of symbols composed of different numbers of vesicle equivalents[21]. The role that MVR might play in coding sensory information has been unclear because it has not yet been observed in response to light or sound.

To understand the functional role of MVR in the retina we used the fluorescent glutamate reporter iGluSnFR[22] to image the synaptic output of bipolar cells in larval zebrafish. Optimizing the signal-to-noise ratio allowed us to count vesicles released from individual active zones *in vivo* and investigate how they are used to represent a stimulus. We find that these ribbon synapses employ a hybrid strategy for encoding visual contrast that involves changes in both the frequency of release events and the amplitude of those events. We term these two components the rate and amplitude codes. Changes in the amplitude code confer several advantages, including the ability to signal contrast beyond the range where the rate code has reached a saturating frequency, improving the temporal precision of transmission and increasing the efficiency of communication by transmitting more bits of information per vesicle.

Results

Optical detection of multivesicular release *in vivo*

All visual information flows through bipolar cells which in turn drive the ganglion cells that deliver the results of retinal processing to other parts of the brain. To investigate how bipolar cells transmit this information across the synapse we drove expression of iGluSnFR using

the *ribeye* promoter[23] (Fig. 1a). A periodic and spatially uniform stimulus (5 Hz) triggered a train of glutamate transients which were sampled by performing line scans through the terminal at 1 kHz (Fig. 1b and c; Supplementary Videos 1 and 2). These signals had Gaussian intensity profiles along the linescan constrained to distances of ~1.5 mm, making it possible to distinguish fusion events at neighbouring active zones by spatial demixing (Fig. 1c and Supplementary Figs. 1 and 2). Glutamate release at adjacent active zones did not always coincide in time, reflecting the stochastic nature of vesicle fusion (Fig. 1d). Most notably, the size of glutamate transients varied widely, both within and between active zones providing the output of one bipolar cell.

Figure 1 near here

The time-course of iGluSnFR transients were similar for events of different amplitude (Fig. 2a), a property that allowed us to improve event detection by use of Wiener deconvolution[6]. The steps in the analysis procedure are described in detail in Online Methods and are summarized in Fig. 2b which shows an example of a raw iGluSnFR trace and its deconvolution using a temporal kernel with the shape shown in Fig. 2a. This approach significantly improved the signal-to-noise ratio (SNR) but it was still necessary to set a threshold to distinguish events from noise, and this was chosen to be at least 3 sd above the baseline (red line in Fig. 2b; detailed examples in Supplementary Figs. 4 and 5). After recording responses to stimuli lasting tens of seconds we constructed histograms of event amplitudes for individual active zones and these displayed several peaks, as shown in Fig. 2c. These peaks were spaced at equal intervals indicating that transients of different sizes were composed of multiples of an underlying unit of transmission, very likely corresponding to a single vesicle (see also Supplementary Fig. 5). Based on these amplitude distributions, we partitioned events of different amplitudes into numbers of quanta using a Gaussian Mixture Model (Fig. 2b, lower trace).

Figure 2 near here

Does the iGluSnFR reporter provide a linear read-out of glutamate release? Evidence that it does is provided in Fig. 2d which plots the relative distance between the peaks in amplitude histograms of the type shown in Fig. 2c. These measurements were collected from six active zones in which large numbers of events of different size were recorded by applying stimuli of contrasts between 20% and 100%. The interpeak distances were not significantly different up to the ninth peak - the largest number that could be

detected routinely within a distribution. These results indicate that the iGluSnFR signal provided a linear read-out of MVR up to at least nine vesicles and that these estimates were not skewed by saturation of the reporter. In comparison, the largest events we observed from a sample of 51 synapses were equivalent to 11 quanta (Fig. 6a).

The electrical signals that generate release events of different amplitude remain to be established but are likely to reflect the fact that the synaptic terminal is not a passive electrical compartment. Regenerative depolarizations can be generated within bipolar cell terminals by the same voltage-sensitive calcium channels that open to trigger vesicle fusion[24] and calcium spikes can phase-lock to visual stimuli with millisecond precision[24-26].

Multiquantal events were a ubiquitous property of bipolar cells in both the ON and OFF channels of the retina (74 ON synapses and 112 OFF). We could not, however, clearly differentiate between strict CMVR (defined as events synchronized within a 100 μs time window) and a transient and very large increase in release probability (P_r) where vesicles are nonetheless released independently (MVR). The best SNR achievable with iGluSnFR only allowed us to distinguish events occurring at intervals greater than ~10 ms (Supplementary Figs. 6-8). Nonetheless, multiple vesicles released in a time-window of 100 μs or 10 ms will summate similarly on a postsynaptic ganglion cell because these have membrane time-constants in the range of 10-40 ms[27-29]. Glutamate release events of different amplitude can therefore be considered as different synaptic symbols (Fig. 1c and d; Fig. 2b).

Reverse-correlation at an active zone: the transmitter-triggered average

Do synaptic events of varying amplitude play a role in encoding a visual stimulus? We began investigating this question by adapting an approach that has been widely used to explore the information represented by spikes – calculation of the spike-triggered average (STA). The STA is obtained by reverse-correlating the spike train generated by a random ("white noise") stimulus to the stimulus itself[2] and provides an estimate of the neurons tuning in the form of its linear filter in time and/or space. By reverse-correlating events measured using iGluSnFR in response to full-field noise we instead constructed the "transmitter triggered average" (TTA) to describe the output of an individual active zone. This approach to understanding the information encoded by vesicles at a synapse was first proposed in 2004[3], but has not been realized until now. There is, however, a fundamental distinction between the STA and TTA: while the spike output of a neuron contains just one symbol, in the case of bipolar cells we are dealing with a vesicle code involving a number of symbols

composed of one, two, three quanta etc. We therefore calculated the TTA separately for synaptic events composed of different numbers of vesicles (Fig. 3).

The TTA revealed that the more quanta within an event the higher, on average, the temporal contrast driving it (Fig. 3b). The relationship between the number of quanta released per event (Q_e) and the contrast in the temporal filter (C) could be described by a first-order saturation of the form $C = C_{max} \times Q_e / (Q_e + Q_{1/2})$, where C_{max} = 19% and $Q_{1/2}$ = 2.4 vesicles (Fig. 3c). Thus, MVR encodes one of the most fundamental properties of a visual stimulus – temporal contrast – in a direct way.

Figure 3 near here

The TTA revealed a second distinction between events of different amplitude: uniquantal events were characterized by temporal filters that were monophasic through both ON and OFF channels, corresponding to a low-pass filter, but the TTAs from ON synapses were biphasic corresponding to band-pass characteristics with peak transmission at ~4Hz (Fig. 3b). In other words, a period of lower light intensity immediately preceding higher intensity favoured larger synaptic events through the ON channel. Such antagonism in the time-domain is commonly observed in the temporal receptive fields of neurons early in the visual system[30], where it has been proposed to underly the suppression of redundant signals[31]. Synchronizing the release of multiple vesicles is therefore expected to amplify the output of a bandpass filter to enhance the signaling of positive temporal contrast.

A hybrid rate and amplitude code

Given that bipolar cells can transmit visual information as changes in both the frequency and amplitude of release events, what are the relative contributions of these two coding strategies? To make such a comparison across synapses it was important to take into account variations in the contrast-response function and compare the modulation of rate and amplitude around similar parts of each synapses operating range. To achieve this, we proceeded in two steps. First, we constructed the contrast-response function of individual active zones using a 5 Hz stimulus in which contrast was increased in steps of 10% (Fig. 4a). The first measure of response that we used was the total number of quanta per cycle of the stimulus (Q_c), from which we estimated the contrast generating the half-maximal response, $C_{1/2}$ (Fig. 4b). The parameter $C_{1/2}$ occurs at the steepest part of the contrast-response function and therefore also defines the range over which the synapse signaled a change in contrast with the highest sensitivity[2]. The second step was to make more detailed measurements of responses in a stimulus range of ±10% around $C_{1/2}$, for which different

contrasts were applied in a pseudo-random order (Fig. 4c). Each contrast was applied for 2 s, again at 5 Hz, and keeping the mean luminance constant throughout.

Figure 4 near here

A first inspection of the responses in Fig. 4a and c indicates that, although the amplitude of events fluctuated throughout each 2 s exposure to a contrast step, higher contrasts tended to increase both the frequency and amplitude of glutamatergic events. To assess how strongly these two components of the synaptic code were modulated Q_c was factorized into the number of events per cycle (E_c, representing the contribution of the rate code) and the number of quanta per event (Q_e, the amplitude code). The relative change in E_c and Q_e varied across our sample of 55 active zones and this could be illustrated by splitting the population according to a simple criterion: is the relative change in E_c more or less than Q_e? In 38 of 55 active zones, an increase in contrast caused E_c to rise more steeply than Q_e and the average behaviour of this subset is shown in Fig. 4d. In the remaining 17 synapses the amplitude code was dominant and Q_e rose more rapidly than E_c (Fig. 4e). This functional hetreogeneity may reflect the fact that different bipolar cells vary in their intrinsic electrophysiological properties as well as the retinal microcircuits in which they are embedded.

Active zones in which the amplitude of synaptic events was modulated more strongly displayed a second notable property: when the average frequency of events reached a maximum (~15 Hz), further increases in contrast were represented wholly as increases in the number of quanta released per event (arrowed in Fig. 4e). MVR therefore extended the range of contrasts that could be signaled beyond those allowed by the rate code alone.

The amplitude code improves the temporal resolution of synaptic transmission

Synaptic events of larger amplitude improved the temporal precision with which the visual signal was transmitted. An example of this phenomenon is shown in Fig. 5a which compares the timing of events relative to the phase of a 5 Hz stimulus for contrasts of 20% and 100%. Uniquantal events displayed a standard deviation ("temporal jitter") of 24 to 28 ms over a range of contrasts, while events composed of 7 or more quanta jittered by as little as 2.5 ms (Fig. 5b). The degree to which synaptic events were consistent in time also depended on the contrast of the stimulus eliciting the event. For instance, the largest packets of glutamate consistently observed at 20% contrast contained 8 quanta with a jitter of 12.1 ± 1.4 ms, while 8 quantal events at 100% contrast jittered by just 4.6 ± 0.6 ms (Fig. 5b). The temporal precision of the largest events was similar to that displayed by the spike

trains in post-synaptic ganglion cells, which also show precisions down to a few milliseconds when responding to high contrast[32]. MVR may therefore be one of the mechanisms that cause the spike output of ganglion cells to become more reliable as contrast is increased.

Figure 5 near here

Multivesicular events transmit more information per vesicle

The responses of sensory neurons can be highly variable, but they nonetheless provide information because the distribution of possible responses alters for different stimuli[2,33]. To begin exploring how synapses of bipolar cells provided information through MVR we measured the change in the distribution of synaptic event amplitudes at high and low contrasts (20% and 100%). Figure 6a shows that unitary events predominated at low contrast, while an event composed of two quanta was equally likely whether the contrast was 20% or 100%. At the other extreme, events composed of more than 8 quanta indicated a contrast above 20% with a very high degree of certainty. The largest events observed at 100% contrast were composed of 11 quanta.

Figure 6 near here

A more systematic investigation of changes in the distribution of events was made by applying information theory[33,34]. Using the stimulus protocol shown in Fig. 4c we quantified the mutual information between a set of stimuli of varying contrasts and release events containing different numbers of quanta. Vesicles released individually carried an average of 0.125 bits of information which is, as expected, significantly less than the 1-3.6 bits transmitted per spike in post-synaptic ganglion cells[32,35]. Larger events transmitted progressively more information (Fig. 6b) because they were rarer overall and correlated with higher contrasts rather than occurring randomly (Figs. 3-5). The relation between the specific information (i, bits) and Q_e, the number of vesicles comprising the event, could be described as a power function with an exponent of ~3, indicating that larger synaptic events transmitted more information per vesicle. To quantify this idea more directly we divided the amount of information in an event by the number of vesicles it contained to provide a measurement of "vesicle efficiency", after which we averaged across synapses to quantify the trend. The change in this quantity is plotted in Fig. 6c after normalizing to the value measured in the same synapse for one vesicle. Synaptic events composed of five vesicles carried, on average, four times as much information per vesicle compared to unitary events.

255

Discussion

Achieving single-vesicle resolution with iGluSnFR has provided the opportunity to investigate the synaptic code by counting quanta released at individual active zones, much as electrophysiology allows the counting of spikes to investigate the neural code[1,2]. This *in vivo* approach demonstrates that the release of two or more vesicles within a time-window of 10 ms or less is a fundamental aspect of the strategy by which ribbon synapses of bipolar cells recode an analogue signal for transmission across the synapse. MVR discretizes the graded signal arriving down the axon into about eleven output values, with larger modulations of intensity increasing the average number of vesicles released within an event (Fig. 6a). As a result, these synapses implement a hybrid coding strategy that represents a stimulus as changes in both the rate and amplitude of synaptic events (Figs. 3-4). The amplitude code complements the rate code by increasing the temporal accuracy (Fig. 5) and operating range (Fig. 4) of synapses transmitting the visual signal to RGCs, as well as the efficiency of information transmission measured as bits per vesicle (Fig. 6).

An amplitude code at ribbon synapses

Neurons in the brain can be characterized as "leaky integrators" in which inputs converging on the dendrites summate increasingly less effectively the further apart they arrive in time and space[2]. The time-window for effective summation is set by the membrane time-constant, which varies between 10-40 ms in retinal ganglion cells[24-26]. In this context, MVR can be viewed as a mechanism by which glutamatergic vesicles are summed *pre-synaptically* and at *one site* over a time-window of one time-constant or less, causing them to act more effectively on the post-synaptic neuron.

One factor determining how efficiently presynaptic summation of glutamate converts into a postsynaptic current is whether or not post-synaptic receptors saturate. Saturation does not appear to occur at AII amacrine cells postsynaptic to rod-driven bipolar cells[18] or auditory nerve fibers postsynaptic to hair cells[20] but may occur at AMPA receptors on RGCs[36]. On the other hand, multivesicular events activate extrasynaptic NMDARs on ganglion cell dendrites much more effectively than single vesicles[37]. We therefore require experimental measurements that explore how varying numbers of vesicles released within an event alter the post-synaptic cell's input current and probability of spiking. Such measurements will be difficult using electrophysiology alone because a patch pipette on the cell body cannot easily distinguish true MVR from the coincidental arrival of vesicles released from different synapses. A direct and reliable assessment of the postsynaptic effects of MVR has, however, been possible in the auditory system of bullfrog where afferents contacting only one hair cell ribbon can be patched. Measuring the synaptic

currents injected into the afferent relative to the currents needed to depolarize the fiber beyond threshold demonstrates that larger glutamatergic events will reliably trigger spikes[20].

Cellular mechanisms underlying multivesicular release

The cellular mechanisms synchronizing the release of two or more vesicles remain to be discovered. It is, however, notable that the synaptic terminals of bipolar cells in zebrafish[27,28] and mice[29] can convert the analogue signal arriving down the axon into regenerative calcium spikes that cause large increases in pre-synaptic calcium. It is also known that strong activation of pre-synaptic calcium channels can release a large fraction of the readily-releasable pool (RRP) of vesicles within a few milliseconds[13-15]. For instance, the RRP at an individual active zone has been estimated to be about 18 vesicles in goldfish bipolar cells[38], indicating that MVR events ranging up to eleven could be generated by the fast release of vesicles that are already docked and primed. Variations in the degree to which the rate and amplitude codes were modulated in different synapses (Fig. 4) might reflect intrinsic factors, such as the complement of ion channels in the terminal membrane or the size of the RRP, or the extrinsic influence of the microcircuit in which the bipolar cell is embedded, including inhibition from amacrine cells.

The ionic mechanisms that generate neural signals and the synaptic processes that transmit them are a major energetic cost to the brain[39,40] and the evolutionary pressure exerted by the need to transmit information in an energy-efficient manner has provided a general principle by which to understand the design of sensory circuits[12,41,42]. This principle provides an explanation for the use of analogue signaling to transmit early visual and auditory information: graded changes in membrane potential transmit information more efficiently than spikes[12,43]. It would be interesting to carry out an energy budget for MVR at synapses of bipolar cells where the efficiency of transmission could be quantified as bits per unit energy [44,45], but this will require better understanding of the underlying electrical events.

The time-resolution of measurements with iGluSnFR do not rule out the possibility that the release of multiple vesicles is synchronised to sub-millisecond time-scales by co-ordinated multivesicular release. The mechanisms that might underlie CMVR are far from clear[17-19], but appear distinct from more common mechanisms of synaptic transmission by deviating from a Poisson process in which vesicles fuse independently of each other. A third suggestion, based on experiments on ribbon synapses of hair cells, has been that variable amounts of glutamate are released from a single vesicle because of the dynamics of a fusion pore[46]. This idea runs counter to evidence that bipolar cells transmit by full collapse of vesicles into the membrane surface[47,48] and the clear quantization of glutamate transients that we observe with iGluSnFR (Fig. 2).

The presence of ribbon structures at synapses driven by analogue signals correlates with a second functional specialization - a continuous mode of operation that allows the transmission of sensory information to be maintained over prolonged periods[9,49] (Figs. 4 and 5). The ribbon is thought to support continuous release by capturing vesicles from a mobile pool in the cytoplasm[38] and then transporting them to the active zone[19]. The maximum rates of continuous release that we observed using iGluSnFR were in the range of 50-100 vesicles s^{-1} per active zone, which is in agreement with measurements made using the membrane dye FM1-43[14,49] and the fluorescent reporter protein sypHy[23]. The efficient replenishment of vesicles within the RRP is an essential aspect of ribbon synapse operation in the retina and allows MVR to continue to operate over prolonged periods, as shown in Figs. 3-5.

Amplitude and rate coding in relation to the output of the retina

For vision to be useful, information about important stimuli must be transmitted over an appropriately short time window. Multiquantal events were rare and strongly dependent on temporal contrast, so their arrival provided more information about a preceding stimulus than vesicles released individually (Fig. 6). Crucially, MVR also encoded the timing of a stimulus more precisely: the largest glutamatergic events jittered by just a few milliseconds relative to a stimulus (Fig. 5), which is similar to the spike responses observed in postsynaptic ganglion cells. The spike trains generated by naturalistic stimuli often consist of brief increases in firing frequency from longer background periods of silence, making it difficult to describe activity as a time-varying rate. Rather, it has been suggested that "firing events containing single spikes or bursts of spikes are elicited precisely enough to convey distinct packets of visual information, and hence may constitute the fundamental symbols in the neural code of the retina"[32]. It seems possible that these symbols originate in the amplitude code of bipolar cell synapses.

To understand why a coding strategy based on amplitude might have arisen, it is useful to think about the temporal requirements of a simple rate code. If a Poisson synapse releasing all vesicles independently encodes an event by changing the rate of vesicle release by a factor k (from R to kR), the signal-to-noise ratio achieved over an observation time Dt will be

$$SNR = \frac{(kR\Delta t - R\Delta t)}{\sqrt{(kR\Delta t + R\Delta t)}}$$

(Online Methods, Equation 4). A change in contrast from $C_{1/2}$ -10% to $C_{1/2}$ +10% increased the rate of vesicle release by an average factor k =1.8 (Q_e x E_c; Fig. 4d), from a basal rate R of no more than 20 vesicles s^{-1}. To detect such a change with a SNR of 3 (a fairly good reliability) would require an observation time of ~2 s. In comparison, any release event with amplitude greater than 8 quanta would signal an increase in contrast beyond 20% within milliseconds (Fig. 6a). This simple comparison illustrates one of the potential advantages of recoding an analogue signal using symbols varying in amplitude rather than by varying the frequency of digital events: an unexpected symbol immediately imparts new information, while a stochastic rate code must be observed over a time window sufficiently long to establish a significant change relative to the background. Further analysis of the statistics of MVR will likely shed more light on the properties of the vesicle code transmitting visual information.

Acknowledgments

Many thanks to J Johnston for discussions and to H Smulders and N Bashford for looking after our zebrafish. Thanks to M Meyer for gifting constructs. This work was supported by grants to LL from the Wellcome Trust (102905/Z/13/Z) and an EU International Training Network (H2020-MSCA-ITN-2015-674901).

Author contributions

BJ wrote software, conceived and designed experiments and helped prepare the manuscript; LD performed molecular biology, and carried out experiments and analysis; JM-D carried out experiments and analysis and helped prepare the manuscript; S-HS performed molecular biology, fish transgenesis and initial functional analysis; LL conceived the project, designed experiments, wrote software, analyzed results and prepared the manuscript.

Competing interests

The authors declare no competing interests.

Figure Legends

Figure 1. Glutamate transients of varying amplitude imaged at individual active zones **a)** Multiphoton section through the eye of a zebrafish larva (7dpf) expressing iGluSnFr in a subset of bipolar cells. **b)** Linescan through a single terminal. No other terminals were in the vicinity. **c)** The kymograph (top) shows the intensity profile along the line in b as a function of time. The broken green trace to the side of the kymograph shows the time-averaged intensity along the line and the red and black traces are the two Gaussians that sum to best describe this spatial profile. The amplitude of each Gaussian at each time point was used to quantify the signal at each active zone, and is plotted in the traces below. The modulation in light intensity (20% contrast, 5 Hz) is shown immediately below. Note the large variations in the amplitude of glutamate transients. **d)** Expansion of the records within the blue dashed boxes in c. Sometimes both active zones release glutamate while on other occasions only active zones 1 or 2 are active (red dotted line). Qualitatively similar iGluSnFR signals were observed in 150 independent experiments.

Figure 2. Glutamatergic events of different amplitudes were composed of varying numbers of quanta. a) The average of 15 iGluSnFr events with a peak amplitude of $\Delta F/F$ = 0.198 ± 0.001 (black; mean ± sd) superimposed on the average of 15 events approximately four times as large (red; peak $\Delta F/F$ = 0.73 ± 0.01). After normalization, the small and large events superimpose. The decline from the peak occurs with a time-constant of 44 ms (dashed line). **b)** An example of the basic steps in the analysis by which events were counted and quantified. The raw iGluSnFR signal is shown at top (stimulus at 5 Hz, 80% contrast; Savitsky-Golay filter 21 ms). The middle trace shows the results of Wiener deconvolution using a kernel of unitary area and the shape shown in **a.** The time and amplitude of each event was obtained from the local maxima above a threshold (dashed red line). Similar analysis could be carried out in 150 independent experiments. **c)** A histogram of event amplitudes for the active zone featured in **b** (n = 547 events accumulated using stimulus contrasts of 30%, 80% and 100%). The black line is a fit of nine Gaussians, identified using a Gaussian Mixture Model. Note that the variance of successive Gaussians did not increase in proportion to the peak number. The first peak had a value of 0.24 and the distance between peaks averaged 0.26, indicating the existence of a quantal event equivalent to ~0.25. The amplitude of the quantal event averaged 0.23 ± 0.01 (mean ± sem, n = 20 synapses from independent experiments). **d)** The mean distance between successive peaks (normalized to the distance between the first and second peaks) plotted as a function of the peak number. Collected results from n = 6 synapses. Points show mean

± sem. The dashed line is a linear fit with a slope of 0.03 ± 0.12 (mean ± sem), which is not significantly different from zero. There were no signs of saturation of the iGluSnFR signal for events composed of up to 9 quanta.

Figure 3. The transmitter-triggered average (TTA) depends on the number of quanta in a release event

a) Example of iGluSnFR signals (top) elicited by a "white noise" stimulus (bottom). Qualitatively similar responses were elicited in 17 independent experiments. **b)** Upper traces: Linear filters extracted by reverse-correlation of responses composed of one quantum (left) and four to six quanta (right) from ON synapses (green, n = 9) and OFF synapses (red, n = 8). Multiquantal events encoded larger modulations in intensity. The plots below show the power spectra of the linear filters (each point is mean ± sem). Through the OFF channel, both uniquantal and multiquantal events were driven through low-pass filters, but in the ON channel multiquantal events were driven through a band-pass filter peaking at ~4 Hz. **c)** Relation between the Michelson contrast represented in the TTA and the number of quanta released within the events used for reverse correlation in ON and OFF terminals (n = 9 and 8, respectively; each point is mean ± sem). The relation could be described by a first-order saturation of the form $C = (C_{max})N/(N + N_{1/2})$, where C_{max} = 19% and $N_{1/2}$ = 2.4.

Figure 4. The relative contributions of coding by rate and amplitude

a) Example of the iGluSnFR signal (top) elicited by a full-field stimulus of increasing contrast delivered at 5 Hz (bottom). This protocol was used to quickly assess the half-point of the contrast-response function ($C_{1/2}$) in 38 independent experiments. **b)** Contrast-response function extracted from the synapse in **a**, with response (R) quantified as the total number of quanta per cycle of the stimulus. The curve is a Hill equation of the form $C = (R_{max})(C^h/C^h + C^h_{1/2})$, where R_{max} = 9.8 quanta per cycle (49 quanta s^{-1}), h = 4.5 and $C_{1/2}$ = 39% (dashed arrow). Each point shows mean ± sem for n = 10 cycles of the stimulus. **c)** A stimulus set for quantifying the modulation of the rate and amplitude of synaptic events in 38 independent experiments. Stimuli of varying contrasts were selected to span ± 10% of the range around which the contrast sensitivity was highest in steps of 2%. Each stimulus lasted 2 s and all were at the same mean luminance and a frequency of 5 Hz. Note variations in the frequency and amplitude of events at different contrast levels. Stimuli were applied in two different sequences which were repeated alternately. The same stimulus set was used to estimate mutual information (Figure 6). **d)** The relative change in synaptic activity around $C_{1/2}$. The average number of events per cycle (E_c, black) is compared with the average number of

quanta per event (Q_e, red). Stimuli were delivered in 2 s episodes with 2 s rest as shown in Fig. 4a. In these n = 38 synapses an increase in contrast caused E_c to rise more steeply than Q_e. **e)** In the remaining n = 17 synapses, Q_e rose more steeply than E_c, which then saturated (dashed line) such that further increases in contrast were signaled only by increases in the number of quanta per event (vertical arrow). Points in **d** and **e** show mean ± sem.

Figure 5. Multivesicular events increased the temporal precision of synaptic transmission

a) Top: iGluSnFR signals from a terminal stimulated at 100% and 20% contrast. Note the large variation in event amplitudes at lower contrast. Bottom: expanded time-scale showing responses at 20% and 100% contrast superimposed on the phase of the stimulus (grey). Events composed of fewer quanta are less synchronized than multiquantal events. Two examples of events occurring at phases different to the majority are shown by the blue arrows**.** Qualitatively similar responses were observed in 70 independent experiments. **b)** Temporal jitter (s) as a function of the number of quanta per event (Qe) at contrasts of 20% (blue, n= 57 synapses), 50% (red, n= 61) and 100% (black, n= 63). Points show mean ± sem. Events composed of more quanta exhibit lower jitter at a given contrast and events of a given amplitude exhibit lower jitter at higher contrasts. The relationship between s and Q_e could be described as $s = s_{min} + (s_{max} - s_{min})(e^{-Qe/k})$. At 100% contrast, s_{min} = 2.7 ms (dashed black line), s_{max} = 30 ms and k = 2.6 quanta. At 20% contrast, s_{min} = 9.5 ms (dashed blue line), s_{max} = 28 ms and k = 2.4 quanta.

Figure 6. Multivesicular events increased the efficiency of the vesicle code

a) Changes in the distribution of event amplitudes elicited by low (20%) and high (100%) contrast. Events composed of more than 8 quanta were not observed at 20% contrast. The largest events observed at 100% contrast were composed of about 11 quanta. **b)** Specific information per event (i, bits) as a function of Q_e, the number of vesicles comprising the event. Each point shows mean ± sem and the curve describing the points is a power function of the form $i = i_o + A.Q_e^x$, with i_o = 0.12 bits, A= 0.008, x = 2.8. Results pooled from n= 17 synapses. **c)** Specific information per vesicle normalized to the value measured for a uniquantal event (i'). Each point shows mean ± sem and the curve describing the points is a power function of exponent 2.1. Results from the same 17 synapses in **b**.

504
505

References

1 Rieke, F., Bialek, W., Warland, D. & Ruyter van Steveninck, R. *Spikes: Exploring the Neural Code*. (The MIT Press, 1999).

2 Dayan, P. & Abbott, L. R. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. (The MIT Press 2005).

3 Abbott, L. F. & Regehr, W. G. Synaptic computation. *Nature* **431**, 796-803, doi:10.1038/nature03010 (2004).

4 Del Castillo, J. & Katz, B. Quantal components of the end-plate potential. *The Journal of Physiology Physiology* **124**, 560-573 (1954).

5 Choi, S. Y. *et al.* Encoding light intensity by the cone photoreceptor synapse. *Neuron* **48**, 555-562 (2005).

6 Malagon, G., Miki, T., Llano, I., Neher, E. & Marty, A. Counting Vesicular Release Events Reveals Binomial Release Statistics at Single Glutamatergic Synapses. *The Journal of Neuroscience* **36**, 4010-4025 (2016).

7 Masland, Richard H. The Neuronal Organization of the Retina. *Neuron* **76**, 266-280 (2012).

8 Fuchs, P. A. Time and intensity coding at the hair cell's ribbon synapse. *The Journal of Physiology* **566**, 7-12, doi:10.1113/jphysiol.2004.082214 (2005).

9 Lagnado, L. & Schmitz, F. Ribbon Synapses and Visual Processing in the Retina. *Annual Review of Vision Science* **1**, 235-262 (2015).

10 Jackman, S. L. *et al.* Role of the synaptic ribbon in transmitting the cone light response. *Nature Neuroscience* **12**, 303-310, doi:10.1038/nn.2267 (2009).

11 Freed, M. A. Quantal encoding of information in a retinal ganglion cell. *Journal of Neurophysiology* **94**, 1048-1056, doi:10.1152/jn.01276.2004 (2005).

12 Sterling, P. & Laughlin, S. B. *Principles of Neural Design*. (MIT Press, 2015).

13 Mennerick, S. & Matthews, G. Ultrafast exocytosis elicited by calcium current in synaptic terminals of retinal bipolar neurons. *Neuron* **17**, 1241-1249 (1996).

14 Neves, G. & Lagnado, L. The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells. *The Journal of Physiology* **515 (Pt 1)**, 181-202 (1999).

15 Burrone, J. & Lagnado, L. Synaptic depression and the kinetics of exocytosis in retinal bipolar cells. *The Journal of Neuroscience***20**, 568-578 (2000).

16 Rudolph, S., Tsai, M. C., von Gersdorff, H. & Wadiche, J. I. The ubiquitous nature of multivesicular release. *Trends Neurosci* **38**, 428-438, doi:10.1016/j.tins.2015.05.008 (2015).

17 Glowatzki, E. & Fuchs, P. A. Transmitter release at the hair cell ribbon synapse. *Nature Neuroscience* **5**, 147-154 (2002).

18 Singer, J. H., Lassova, L., Vardi, N. & Diamond, J. S. Coordinated multivesicular release at a mammalian ribbon synapse. *Nature Neuroscience* **7**, 826-833 (2004).

19 Mehta, B., Snellman, J., Chen, S., Li, W. & Zenisek, D. Synaptic ribbons influence the size and frequency of miniature-like evoked postsynaptic currents. *Neuron* **77**, 516-527 (2013).

20 Li, G. L., Cho, S. & von Gersdorff, H. Phase-locking precision is enhanced by multiquantal release at an auditory hair cell ribbon synapse. *Neuron* **83**, 1404-1417 (2014).

21 DeWeese, M. R. & Meister, M. How to measure the information gained from one symbol. *Network: Computation in Neural Systems* **10**, 325-340 (1999).

22 Marvin, J. S. *et al.* An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nature Methods* **10**, 162-170 (2013).

23 Odermatt, B., Nikolaev, A. & Lagnado, L. Encoding of luminance and contrast by linear and nonlinear synapses in the retina. *Neuron* **73**, 758-773 (2012).

24 Dreosti, E., Esposti, F., Baden, T. & Lagnado, L. In vivo evidence that retinal bipolar cells generate spikes modulated by light. *Nature Neuroscience* **14**, 951-952, doi:10.1038/nn.2841 (2011).

25 Baden, T., Esposti, F., Nikolaev, A. & Lagnado, L. Spikes in retinal bipolar cells phase-lock to visual stimuli with millisecond precision. *Current Biology* **21**, 1859-1869 (2011).

26 Baden, T., Berens, P., Bethge, M. & Euler, T. Spikes in mammalian bipolar cells support temporal layering of the inner retina. *Current Biology* **23**, 48-52, doi:10.1016/j.cub.2012.11.006 (2013).

27 Taylor, W. R., Mittman, S. & Copenhagen, D. R. Passive electrical cable properties and synaptic excitation of tiger salamander retinal ganglion cells. *Vis Neurosci* **13**, 979-990 (1996).

28 Robinson, D. W. & Chalupa, L. M. The intrinsic temporal properties of alpha and beta retinal ganglion cells are equivalent. *Current Biology* **7**, 366-374 (1997).

29 O'Brien, B. J., Isayama, T., Richardson, R. & Berson, D. M. Intrinsic physiological properties of cat retinal ganglion cells. *The Journal of Physiology* **538**, 787-802 (2002).

30 Meister, M. & Berry, M. J., 2nd. The neural code of the retina. *Neuron* **22**, 435-450 (1999).

31 Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **216**, 427-459 (1982).

32 Berry, M. J., Warland, D. K. & Meister, M. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5411-5416 (1997).

33 Stone, J. V. *Principles of Neural Information Theory: Computational Neuroscience and Metabolic Efficiency* (Sebtel Press, 2018).

34 Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and Information in Neural Spike Trains. *Physical Review Letters* **80**, 197-200 (1998).

35 Koch, K. *et al.* How Much the Eye Tells the Brain. *Current Biology* **16**, 1428-1434, (2006).

36 Sagdullaev, B. T., McCall, M. A. & Lukasiewicz, P. D. Presynaptic inhibition modulates spillover, creating distinct dynamic response ranges of sensory output. *Neuron* **50**, 923-935 (2006).

37 Chen, S. & Diamond, J. S. Synaptically released glutamate activates extrasynaptic NMDA receptors on cells in the ganglion cell layer of rat retina. *The Journal of Neuroscience* **22**, 2165-2173 (2002).

38 Holt, M., Cooke, A., Neef, A. & Lagnado, L. High mobility of vesicles supports continuous exocytosis at a ribbon synapse. *Current Biology* **14**, 173-183 (2004).

39 Laughlin, S. B., de Ruyter van Steveninck, R. R. & Anderson, J. C. The metabolic cost of neural information. *Nature Neuroscience* **1**, 36-41 (1998).

40 Attwell, D. & Laughlin, S. B. An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab* **21**, 1133-1145 (2001).

41 Barlow, H. B. in *Sensory Communication* (ed W. A. Rosenblith) 217-234 (MIT Press, 1961).

42 Niven, J. E. & Laughlin, S. B. Energy limitation as a selective pressure on the evolution of sensory systems. *J Exp Biol* **211**, 1792-1804 (2008).

43 de Ruyter van Steveninck, R. R. & Laughlin, S. B. The rate of information transfer at graded-potential synapses. *Nature* **379**, 642, doi:10.1038/379642a0 (1996).

44 Harris, J. J., Jolivet, R. & Attwell, D. Synaptic energy use and supply. *Neuron* **75**, 762-777 (2012).

45 Harris, Julia J., Jolivet, R., Engl, E. & Attwell, D. Energy-Efficient Information Transfer by Visual Pathway Synapses. *Current Biology* **25**, 3151-3160 (2015).

46 Chapochnikov, N. M. *et al.* Uniquantal release through a dynamic fusion pore is a candidate mechanism of hair cell exocytosis. *Neuron* **83**, 1389-1403 (2014).

47 Llobet, A., Beaumont, V. & Lagnado, L. Real-time measurement of exocytosis and endocytosis using interference of light. *Neuron* **40**, 1075-1086 (2003).

48 Zenisek, D., Davila, V., Wan, L. & Almers, W. Imaging calcium entry sites and ribbon structures in two presynaptic cells. *The Journal of Neuroscience* **23**, 2538-2548 (2003).

49 Lagnado, L., Gomis, A. & Job, C. Continuous vesicle cycling in the synaptic terminal of retinal bipolar cells. *Neuron* **17**, 957-967 (1996).

50 Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed Eng Online* **2**, 13, (2003).

Methods

Zebrafish husbandry

Fish were raised and maintained under standard conditions on a 14 h light/10 h dark cycle[23]. To aid imaging, fish were heterozygous or homozygous for the casper mutation which results in hypopigmentation and they were additionally treated with1-phenyl-2-thiourea (200 μM final concentration; Sigma) from 10 hours post fertilization (hpf) to reduce pigmentation further. All animal procedures were performed in accordance with the Animal Act 1986 and the UK Home Office guidelines and with the approval of the University of Sussex Animal Welfare and Ethical Review Board. More information about experimental design and reagents is available in the Life Sciences reporting Summary.

Molecular Biology

The zebrafish ribeye a (ctbp2) promoter was used to drive expression of iGluSnFR in all neurons with ribbon synapses. Tg(–1.8ctbp2:Gal4VP16_BH) fish that drive the expression of the transcriptional activator protein Gal4VP16 were generated by co-injection of I-SceI meganuclease and endofree purified plasmid into wild-type zebrafish with a mixed genetic background. A myocardium-specific promoter that drives the expression of mCherry protein was additionally cloned into the plasmid to allow for phenotypical screening of founder fish. Tg(10xUAS:iGluSnFR_MH) fish driving the expression of the glutamate sensor iGluSnFR[22] under the regulatory control of the 10 x UAS enhancer elements were generated by co-injection of purified plasmid and tol2 transposase RNA into offspring of AB wildtype fish outcrossed to casper wildtype fish. The sequences for the myocardium-specific promoter driving the expression of enhanced green fluorescent protein (mossy heart) were added to the plasmid to facilitate the screening process. Supplementary Table 1 shows plasmid and primer information.

Multiphoton Imaging *In Vivo*

Zebrafish larvae (7–9 days post-fertilization) were immobilized in 3% low melting point agarose (Biogene) in E2 medium on a glass coverslip (0 thickness) and mounted in a chamber where they were superfused with E2[23]. Imaging was carried out using a two-photon microscope (Scientifica) equipped with a mode-locked titanium-sapphire laser (Chameleon, Coherent) tuned to 915 nm and an Olympus XLUMPlanFI 20x water immersion objective (NA 0.95). To prevent eye movements, the ocular muscles were paralyzed by injection of 1 nL of α-bungarotoxin (2 mg/mL) behind the eye. The signal-to-noise ratio for imaging was optimized by collecting photons through both the objective and a sub-stage oil condenser (Olympus, NA 1.4). Emission was filtered through GFP filters (HQ 535/50,

Chroma Technology) before detection with GaAsP photomultipliers (H7422P-40, Hamamatsu). The signal from each detector passed through a current-to-voltage converter and then the two signals were added by a summing amplifier before digitization. Scanning and image acquisition were controlled under ScanImage v.3.6 software[50]. Images were typically acquired at 10 Hz (128 × 100 pixels per frame, 1 ms per line) while linescans were acquired at 1 kHz.

Full-field light stimuli were generated by an amber LED (lmax = 590 nm, Thorlabs), filtered through a 590/10 nm BP filter (Thorlabs), and delivered through a light guide placed close to the eye of the fish. Stimuli were normally delivered as modulations around a mean intensity of ~165 nW/mm^2 and the microscope was synchronized to visual stimulation.

To minimize the possibility of recording iGluSnFR signals from adjacent synapses, we used zebrafish in which only a fraction of bipolar cells were expressing iGluSnFR and chose terminals which were spatially isolated from others expressing the reporter, as shown in Fig. 1a and Supplementary Figs. 9 and 10. These transients were generated by the electrical signal arriving from the cell body rather than by glutamate spillover from neighbouring neurons because they were destroyed by ablating the soma of the cell (Supplementary Figure 9).

Counting ribbons

To assess the probability of conflating signals from different active zones within one terminal we measured the numbers and distribution of synaptic ribbons that holds vesicles close to the sites of fusion[12]. Larvae were fixed at 7 dpf with 4% PFA (in PBS) for 35 min. The retina was gently removed leaving behind the pigment epithelium and sliced into two equal parts, which were permeabilized for 10 min in PBS containing 0.5%. The retinal pieces were incubated for 3 days in primary antibodies to ribeye A (1/200 dilution) and GFP (1/500 dilution, targeting iGluSnFR; Torrey Pines Biolabs, TP401, Lot number 42704), followed by incubation overnight in a secondary antibody, all at 4^{o}C on a shaker. The primary antibody against zebrafish Ribeye A [C]-YNQGYLDRPDPRNIRK-[N] was an isolated chicken IgY fragment produced by Cambridge Research Biochemicals (clone 2575) and has been validated in the laboratories of Leon Lagnado, Rachel Wong and Leanne Godinho. The secondary antibody for ribeye A was Alexa-546 anti-chicken (Molecular Probes, Cat. A11040, Lot. 682609, dilution 1:1000) and for GFP we used Alexa-647 anti-rabbit (Molecular Probes, A21244, Lot. A21244, dilution 1:1000). A confocal microscope (Leica SP8) was used to image through the retina using z steps of 0.25 μm and ImageJ was used to make 3D reconstructions of individual terminals using the “3D Blob segmentation” plugin. The number of ribbons in a sample of 27 terminals averaged 4.6 ± 2.4 (mean ± sd). The distance from

one ribbon to its neighbours averaged 0.96 ± 0.4 μm. The probability of conflating signals from two active zones was estimated to be less than 8% (see below).

Statistics

No statistical methods were used to predetermine sample sizes: experiments were repeated until trends in results were clear and this resulted in sample sizes at least equivalent to previous publications[15,18-20]. All data are given as mean ± s.e.m. unless otherwise stated in the figure legends. All statistical tests were calculated using inbuilt functions in IgorPro (Wavemetrics) and met the assumptions of the statistical tests used. When data were not normally distributed we used non-parametric tests. All tests were two-sided and significance defined as $P < 0.05$. Data collection was not randomized because all experiments were carried out within one set of animals. Delivery of different stimuli was randomized where appropriate. Data collection and analysis were not carried out blind to the conditions of the experiments. Data were only excluded from the analysis if the signal-to-noise ratio (SNR) of the iGluSnFR signals elicited at a given synapse was not sufficient to detect unitary responses to visual stimuli with a SNR of at least three.

Calculation of temporal jitter

In order to compute the temporal jitter of the glutamatergic events, we first calculated the vector strength:

$$\boldsymbol{VS_q = \frac{1}{N_q}\sqrt{\left(\sum_{i=1}^{N_q} cos\left(\frac{2\pi t_{q_i}}{T}\right)\right)^2 + \left(\sum_{i=1}^{N_q} sin\left(\frac{2\pi t_{q_i}}{T}\right)\right)^2}} \quad \#(1)$$

where t_{qi} is the time of the i^{th} q-quantal event, T is the stimulus period, and N_q is the total number of events of composed of q-quanta. The temporal jitter can then be computed by:

$$\boldsymbol{TJ_q = \frac{\sqrt{2(1 - VS_q)}}{2\pi f}} \quad \#(2)$$

where f is the stimulus frequency.

The signal-to-noise ratio associated with a change in the rate of a Poisson process

Imagine the mean rate of a Poisson process, R, changes by a factor k. The signal, S, generated by comparing two observation times Dt will be the change in the mean number of

events counted in each period (kR Dt - R Dt), and the variance in that signal will be the sum of the number of events counted in each period (kR Dt + R Dt). Defining the SNR in the same way as the discriminability (d') used in signal detection theory[45], we have

$$SNR = \frac{S}{\sqrt{variance}} \quad (3)$$

yields

$$SNR = \frac{(kR\Delta t - R\Delta t)}{\sqrt{(kR\Delta t + R\Delta t)}} \quad (4)$$

which can be rearranged to obtain the time period Dt required to obtain a given SNR, as

$$\Delta t = \frac{SNR^2}{R} \frac{(k+1)}{(k-1)^2} \quad (5)$$

Calculation of the Transmitter Triggered Average (TTA)

The TTA for an event containing a specific number of quanta (s_q) was calculated by averaging the stimuli that preceded each release event of that type:

$$\overrightarrow{s_q} = \frac{1}{n_q} \sum_{i=1}^{n_q} \vec{s}(t_i) \#(6)$$

where q is the quantal event type, n_q is the number of n-quantal events within the recording, and s(t) is the stimulus window ending at time t. We used Gaussian White Noise (GWN) in which intensities were drawn from a Gaussian distribution updated at a specific stimulus frame rate (varied between 10 and 30 Hz). The GWN was then discretized into eight equally spaced bins to produce a stimulus that approximates a Gaussian distribution while driving synaptic responses effectively.

Calculations based on information theory

To quantify the amount of information about a visual stimulus that is contained within the sequence of release events from an active zone we first needed to convert bipolar cell outputs into a probabilistic framework from which we could evaluate the specific information (I_2), a metric that quantifies how much information about one random variable is conveyed by the observation a specific symbol of another random variable[21]. The time series of quantal events was converted into a probability distribution by dividing into time bins of 20 ms, such that each bin contained either zero events or one event of an integer amplitude.

We then counted the number of bins containing events of amplitude 1, or 2, or 3 etc. By dividing the number of bins of each type by the total number of bins for each different stimulus, we obtained the conditional distribution of **Q** given **S**, $p(\boldsymbol{Q}|\boldsymbol{S})$, where **Q** is the random variable representing the *quanta/bin* and **S** is the random variable representing the *stimulus contrasts* presented throughout the course of the experiment. We then computed the joint probability distribution by the chain rule for probability (given the experimentally defined uniform distribution of stimuli **S**):

$$\boldsymbol{p(S,Q) = p(Q|S)p(S)} \#(\mathbf{7})$$

In order to convert this distribution into the conditional distribution of S given Q, we used the definition of the conditional distribution:

$$\boldsymbol{p(S|Q) = \frac{p(S,Q)}{p(Q)}} \#(\mathbf{8})$$

From these distributions we computed the specific information as the difference between the entropy of the stimulus S minus the conditional entropy of the stimulus given the observed symbol in the response q:

$$\boldsymbol{I_2(S,q) = H(S) - H(S|q)} \#(\mathbf{9})$$

$$\boldsymbol{I_2(S,q)} = -\sum_{\boldsymbol{s\in S}} \boldsymbol{p(s)\,log\,p(s)} + \sum_{\boldsymbol{s\in S}} \boldsymbol{p(s|q)\,log\,p(s|q)} \qquad (\mathbf{10})$$

representing the amount of information observing each quantal event type q ϵ **Q** carries about the stimulus distribution **S**.

Measuring entropy and mutual information from neural responses can be a challenging problem[33,51]. Estimates require sampling from an unknown discrete probability distribution, and in many cases recording sufficient samples to observe all non-zero probability events is neither tractable nor practical. The biases introduced by undersampling can be a particular problem when the full support of the distribution (all values that map to non-zero probabilities) is high. Within the past few decades, various approaches to correcting biases in information theoretic analyses have been developed[51]. However, as the distributions of interest in this work have both a small support and are well sampled, we have opted to use standard estimates for the quantities of interest.

Analysis sequence for the quantal decomposition of iGluSnFR signals

An analysis package was created within Igor Pro (Wavemetrics) to detect and quantize glutamatergic events, comprising six major steps:

1. Separation of ROIs by spatial decomposition
2. Time series extraction by weighted averaging
3. Baseline correction and calculation of ΔF/F
4. Identification of events by Wiener deconvolution
5. Extraction of events
6. Amplitude clustering to create a time series of quantized events

The analysis begins with an ***x*** x ***L*** matrix in which ***x*** is the number of pixels per line and ***L*** is the number of lines, and provides a final output of a set of ***E*** event times and **AQ** estimated quanta for each event in each active zone. An example of the analysis proceeding from steps 1 to 6 is shown in Supplementary Fig. 1.

ROI detection by spatial decomposition

To define ROIs within linescans, we first noted Fick's second law of diffusion describing the spatio-temporal diffusion of a substance:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}, \#(11)$$

where the change in concentration (***C***) can be described in one dimension (***x***) over time (***t***) with a diffusion coefficient **(D).** If ***N*** is the initial number of glutamate molecules released instantaneously at a point (corresponding to a synaptic release site), a solution can be in the form:

$$c(x,t) = \frac{N}{\sqrt{2\pi\sigma^2(t)}} * \exp\left[-\frac{x^2}{2\sigma^2(t)}\right], \#(12)$$

where

$$\sigma^2(t) = 2Dt. \#(13)$$

This function is recognizable as a normal distribution with variance a monotonically increasing function of time. We only sampled each location at intervals of 1 ms, which made an accurate analytical fit to this function unreliable. We therefore measured the spatial profile as a temporal average of the fluorescence signal along the linescan and fit this average to a sum of Gaussians, where each Gaussian component can be considered its own point source corresponding to an active zone. One of these fluorescence profiles is shown in Fig. 2 defining two nearby active zones. Using a GUI, the user specifies the peak(s) in the profile by

placing one or more cursors and we then use IgorPro's built-in curve fitting routines to fit the temporal average to the function k(x):

$$k(x) = \sum_{i=1}^{n} \frac{Ai}{\sqrt{\left(2\pi\sigma_i^2\right)}} * \exp\left[\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right] \#(14)$$

under the constraints

$$\mu_i \in [c_i - \delta, c_i + \delta],$$
$$A_i > 0,$$

where ***c*** is the set of cursor locations, ***μ*** the set of component means, ***A*** the set of component amplitudes, ***δ*** is a small value allowing for errors in user cursor placement, and ***n*** is the number of placed cursors. The potential problem of overfitting the intensity profile with multiple Gaussians was avoided by restricting the number of components to the number of placed cursors (Supplementary Fig. 2). This process could be repeated with different initial estimates of the locations of the peaks (i.e. the number of cursors and their positions) until the error function reaches a threshold or until the fit was acceptable. User input at this stage allowed us to limit the analysis to active zones with distinct peaks and FWHM <1.5 μm, thereby reducing the possibility of conflating signals from multiple active zones (see "Isolating iGluSnFR signals from individual active zones" below).

Once each spatial component had been defined, a time-series for that component, F(t), was computed as the weighted average of the raw linescan matrix with the spatial filter estimated in step 1:

$$F(t) = \sum_{x} F(x,t)k(x)\,, \#(15)$$

where **F(x,t)** is the raw linescan matrix and **k(x)** is the Gaussian component for the ROI. Increasing the weight of pixels located towards the centre of the spatial profile (Supplementary Fig. 2) allows for significant denoising.

Bleaching of iGluSnFR sometimes occurred during an observation episode and was usually corrected using a linear function of time F(t). The iGluSnFR signal used for all analysis was the relative change in fluorescence, $\Delta F/F_0$, calculated from the bleach-corrected signals. The most frequent value (i.e. the baseline) of the trace was used as F_0.

Identification of events by Wiener deconvolution

Release events within an active zone were identified by their characteristic kinetics using a Wiener filter. The time-course of 101 iGluSnFR transients that were clearly separated in time from other events are shown in Supplementary Fig. 3 and could be described by a function of the form:

$$h(t) = A * exp\left[-\frac{t}{\tau_f}\right] * \left(1 - exp\left[-\frac{t}{\tau_r}\right]\right), \#(16)$$

where ***A*** describes the amplitude of the event and τ_r and τ_f and are the time constants for rise and fall in the signal, respectively. We found that transients at most synapses could be accurately described using a kernel with parameters of τ_f = 0.06 s and τ_r = 0.001 s. These parameters were relatively invariant for transients of different amplitudes, indicating that the reporter operated linearly over the range of glutamate concentrations that we observe (Supplementary Fig. 3; see also Fig. 2): iGlUSnFR signals therefore reflected a linear time-invariant system (LTI), fulfilling the assumptions required for the use of Wiener deconvolution. The result of the Wiener deconvolution was a time series in which glutamate release events were described approximately as Dirac-δ impulse functions of varying amplitudes, as shown by traces in Supplementary Fig. 1d.

Extraction of events

Although the use of Wiener deconvolution significantly improved the signal-to-noise ratio (Fig. 2), it was still necessary to set a threshold to distinguish events from noise. A second example analysis highlighting this key step is provided in Supplementary Figs. 4 and 5. Supplementary Fig. 4a shows a kymograph of a line scan through a terminal in which there were two sources of glutamate, (active zones 1 in red and active zone 2 in black), together with the activity time-series for each obtained after spatial demixing. The corresponding traces after Wiener deconvolution are shown in Supplementary Fig. 4b, where it can again be seen that glutamatergic events varied widely in amplitude. The baseline in the deconvolved traces was not, however, noiseless, making it necessary to set a threshold for counting a deviation in this signal as an event. To choose this threshold, we first examined the distribution of values in the deconvolved trace. Across all experiments, these distributions were consistently Gaussians centred at or very close to zero, except for a small tail of positive values. We therefore used the standard deviation of a Gaussian fit to the distribution to set the threshold at which positive values in the deconvolved trace were considered to be significant (i.e to reflect iGluSnFR events). The thresholds we used were 3-4 standard deviations above the baseline, as shown by the dashed blue lines in Supplementary Fig. 4c. Events were then timed at the local maximum in the deconvolved trace above this threshold, as shown by the dashed vertical lines in the expanded traces in Supplementary Fig. 5. The activity within an active zone could then be described by a vector **E** of event times and **A** of event amplitudes.

To provide some examples of how this procedure performed, the parts of the records highlighted by the large dashed green box in Supplementary Fig. 4a are shown expanded in

Supplementary Fig. 5a. It can be seen, for instance, that the Wiener deconvolution allowed us to distinguish four events closely spaced around 43 s in active zone 1 and two overlapping events around 44 s in active zone 2. Positive deviations in the deconvolved trace that did not exceed threshold were not counted as, for instance, around 46 s in active zone 1. Further examples of small slow deviations of the baseline that were not counted as events are shown in Supplementary Fig. 5b, which is an expansion of the activity in active zone 2 over the period shown by the smaller green box in Supplementary Fig. 4a. Of the last three upward deviations in the raw iGluSnFr signal after 29 s, only the second caused a deviation in the deconvolved trace large enough to cross the threshold for significance. Again, the setting of a threshold after deconvolution rejected small, slowly rising, bumps in the record that did not have the same shape as the Wiener kernel obtained by averaging large numbers of individual events (Supplementary Fig. 3). A histogram of event amplitudes extracted from this active zone is shown in Supplementary Fig. 5c. The black trace is a fitted sum of six Gaussians. The quantal amplitude, as defined by the average inter-peak distance (0.21), is very similar to the amplitude of the first peak (0.23), indicating that the first peak in the distribution represents vesicles released individually.

Amplitude clustering and quantal time series

Based on the evidence that glutamate transients of varying amplitude were integer multiples of a unitary event or quantum (Fig. 2), we partitioned events into numbers of quanta using a Gaussian Mixture Model (GMM). Under this framework, the probability **p** of a value **x** is given by:

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}\left(x \middle| \mu_i, \sigma_i^2\right) \#(17)$$

where **N** is the normal probability density function and **Φ** represents the mixing probability and sums to one. Several algorithms for clustering were tested and we found that Expectation-Maximization (EM) provided a quick and efficient approach. For each synapse, the algorithm was run up to 15 times, each with a different number of components and with the mean and variance parameters initialized randomly. The algorithm was iterated until acceptable convergence had been reached (defined by a user-set threshold). Following completion of the 15 runs, the output clusters were plotted with a histogram of the extracted events to allow the user to select the optimal partition based on these outputs, as well as a plot of data likelihood for each run. This defines the vector **AQ** of estimated number of quanta for each event, an example of which is shown in Supplementary Fig. 1d and Fig. 2 of

914 the main text. Defining a time series as the number of quanta within each event allowed for
915 computation of vesicle release rates and information theoretic measures.
916
917 **Assessment of time resolution**
918 The detection of events relied upon finding local maxima in the deconvolved traces. Noise
919 within our records created the potential problem of mislabelling two individual events as a
920 single, higher quantal event, as shown by the simulation in the left panel of Supplementary
921 Fig. 7. We therefore assessed the 'temporal discrimination window' describing the minimum
922 time between events required before events can be reliably discriminated. First we
923 measured the signal-to-noise ratios (SNR) within our experimental data and then simulated a
924 series of pulses of increasing inter-event intervals with matching SNR values. We then ran
925 each of these simulations through the analysis sequence described above to estimate the
926 temporal discrimination window.
927 The SNR within a recording was defined as the average amplitude of a uniquantal
928 event divided by the standard deviation of the baseline noise. To compute the standard
929 deviation of the noise signal we plotted the distribution of fluorescence values of the raw ΔF/F
930 trace, found the first peak and then fit all values to the left of this peak (thus removing any
931 possible contamination by signal) with a Gaussian, as shown in Supplementary Fig. 6. Note
932 that this is not the same trace used to compute the threshold for event detection: that
933 threshold was applied after deconvolution. In a sample of 10 synapses, the SNR estimated in
934 different traces ranged from three to eight, with the large majority greater than four. In 100
935 simulations using SNRs ranging from 3 to 8 (Supplementary Fig. 7), a SNR of four provided
936 a temporal discrimination window of 10-15 ms (Supplementary Fig. 8).
937
938
939 **A test for the possibility of glutamate spillover**
940 Do glutamate transients recorded by iGluSnFR on the surface membrane of a bipolar cell
941 terminal reflect glutamate release from that same cell or might they also reflect glutamate
942 released from nearby cells? To investigate this question, we tested whether iGluSnFR
943 transients required an electrical signal to arrive from the soma. Supplementary Fig. 9 shows
944 responses from the terminals of two nearby bipolar cells. Destroying the soma of cell 1 using
945 the IR laser (800 nm) destroyed the transients in the connected terminal but not in nearby
946 terminal 2 (Supplementary Fig. 9b). In other words, iGluSnFR molecules on terminal 1 did
947 not generate any significant responses to glutamate released from neighbouring terminals
948 but did require electrical drive arriving from the soma.
949
950 **Testing the linearity of the iGluSnFR signal**

Glutamate transients within the synaptic cleft can reach millimolar concentrations within the first hundreds of microseconds after a vesicle fuses[52]. If a substantial fraction of iGluSnFR molecules providing the signal bind glutamate when a single vesicle is released, it might be expected that the relationship between the iGluSnFR signal and the number of vesicles within an event will gradually saturate. The dissociation constant (K_d) of the iGluSnFR variant used in this study is ~4 μm[22] but it is difficult to judge from such an equilibrium measurement what fraction of iGluSnFR molecules on a spatial scale of ~1 μm become occupied during an iGluSnFR transient (Fig. 1, Supplementary Fig. 2). We therefore took an experimental approach to assess whether iGluSnFR transients might begin to saturate in response to larger MVR events. Amplitude histograms from individual active zones were constructed using stimuli of both low contrast, when the distribution of event amplitudes is shifted towards smaller numbers of quanta, and high contrast, when there are more large events (Fig. 5). An example of such a histogram is shown in Fig. 2c, where eight distinct peaks are evident. We then measured the interpeak distances from a sum of Gaussians fit and asked whether the distance between peaks might be reduced for events containing larger numbers of quanta, as would be expected if there was significant saturation of the reporter. Collected results from six active zones are shown in Fig. 2d. The change in interpeak distance was not significantly different from zero up to events composed of 9 quanta, indicating almost perfect linearity over this range. In comparison, the largest events we observed from a sample of 51 synapses were equivalent to 11 quanta (Fig. 5c). It therefore seems unlikely that our estimates of quantal number were skewed by saturation of the iGluSnFR reporter.

Isolating iGluSnFR signals from individual active zones

The point-spread function (psf) of the microscope had a $FWHM_{xy}$ of 0.7 μm in the x-y, allowing active zones separated by 1 μm to be easily distinguished in the iGluSnFR signal along a single linescan (Fig. 1 and Supplementary Fig. 2). The psf in the *z* dimension was, however, significantly larger ($FWHM_z$ = 2.2 μm), raising the possibility that two or more active zones at different *z* depths might be considered as one if they coincided closely enough in the *x* dimension of the linescan. To assess the probability of conflating signals from different active zones we measured the numbers and distribution of synaptic ribbons that holds vesicles close to the active zone[9,53]. Ribbons within individual terminals were visualized in fixed samples using an antibody to ribeye a, as shown in Supplementary Fig. 10a and b. The ImageJ tool DiAna, was used for object-based 3D co-localization and distance analysis[54]. The distance between ribbons was measured between their centres of mass, but we did not count "floating" ribbons, defined as those that were more than 0.5 mm

from the surface membrane (Euclidean distance). Floating ribbons that are not attached to the surface are a common feature of bipolar cells[53] and accounted for 28% of 58 ribbons in a sample of 4 terminals. The density of ribbons attached to the surface membrane averaged 0.16 mm^{-2}, which was similar to a previous measurement of 0.12 ribbons mm^{-2} made in regions of flattened membrane in bipolar cells from goldfish[53]. The distance between nearest ribbons averaged 0.96 ± 0.4 μm, and the distribution of values is shown in Supplementary Fig. 11b.

To compute the probability of 'collapsing' the signal from two separate ribbons, we first calculated the lateral and axial resolutions of our microscope[55] (ω_{xy} and ω_z):

$$\omega_z = \frac{FWHM_z}{2\sqrt{ln2}} = 1.3\ \mu m \#(18)$$

$$\omega_{xy} = \frac{FWHM_{xy}}{2\sqrt{ln2}} = 0.42\ \mu m \#(19)$$

We then constructed a 'resolution' volume by creating an ellipsoid with major axis equal to the axial resolution and minor axes equal to the lateral resolution, given by equation:

$$\frac{x^2}{{\omega_{xy}}^2} + \frac{y^2}{{\omega_{xy}}^2} + \frac{z^2}{{\omega_z}^2} = 1 \#(20)$$

Thus, a ribbon at the origin of this ellipsoid would not be discriminated from any other ribbon lying within this volume. A section through this "resolution volume" is illustrated in Supplementary Fig. 11a, on which is superimposed the "nearest neighbour volume" defined by a second ribbon at a distance "m" from the first, in any direction. Assuming that all ribbons are randomly distributed relative to each other, the distribution of ribbons is uniform over the surface of the sphere defined by radius "m". The probability of collapsing two ribbons at distance "m" can then be calculated as the surface area of the intersection of the two volumes divided by the surface area of the sphere. The resolution ellipsoid defined in equation 20 is a prolate spheroid, so this probability can be computed analytically or numerically using spherical caps. Supplementary Fig. 11b shows the probabilities of collapsing two ribbons as a function of "m". For the sample of 4 terminals and 42 non-floating ribbons, we compute an average probability of collapsing two nearest ribbons as 8%. These measurements indicate that the large majority of multiquantal events originate from single active zones.

Code availability

The code used to analyze the data in this study is available at https://github.com/lagnadoLab/glueSniffer.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

51 Pola, G., Schultz, S. R, Petersen, R. S & Panzeri, S. in *Neuroscience Databases: A Practical Guide* (Springer, Boston, MA, 2003).

52 Budisantoso, T. *et al.* Evaluation of glutamate concentration transient in the synaptic cleft of the rat calyx of Held. *The Journal of Physiology* **591**, 219-239, (2013).

53 Zenisek, D., Horst, N. K., Merrifield, C., Sterling, P. & Matthews, G. Visualizing synaptic ribbons in the living cell. *The Journal of Neuroscience* **4**, 9752-9759 (2004).

54 Gilles, J. F., Dos Santos, M., Boudier, T., Bolte, S. & Heck, N. DiAna, an ImageJ tool for object-based 3D co-localization and distance analysis. *Methods* **115**, 55-64 (2017).

55 Zipfel, W. R., Williams, R. M. & Webb, W. W. Nonlinear magic: multiphoton microscopy in the biosciences. *Nature Biotechnology* **21**, 1369 (2003).

Figure 1

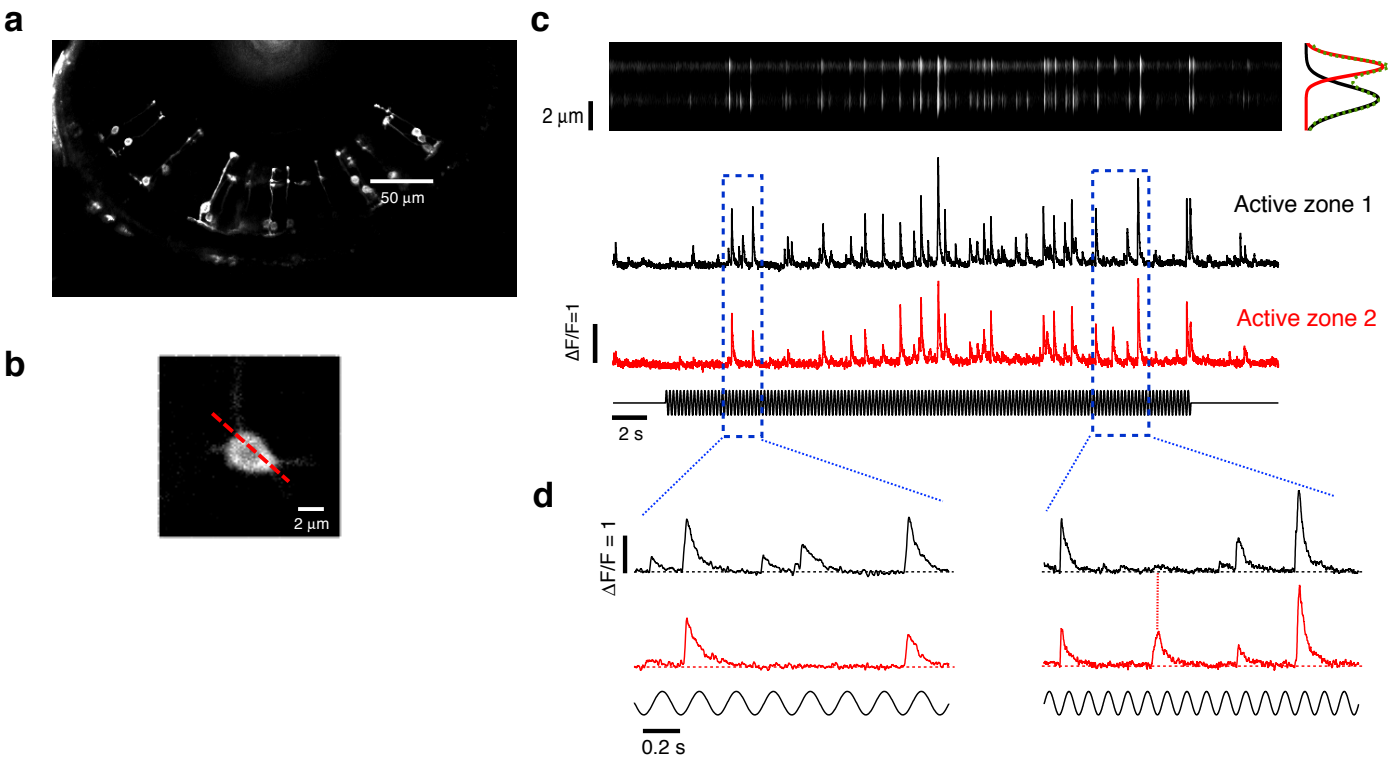


Figure 2

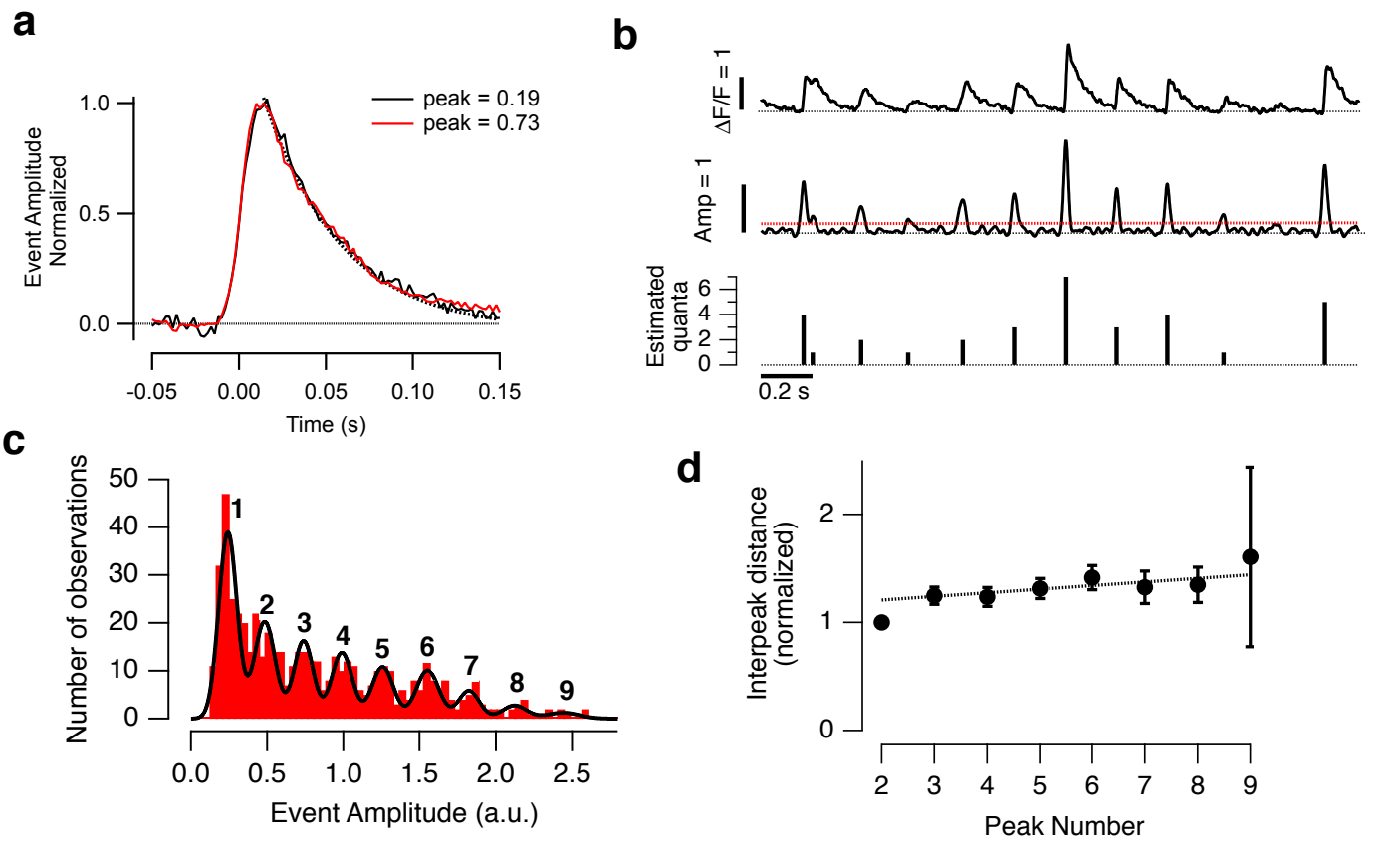

a
Event Amplitude Normalized
1.0
0.5
0.0
peak = 0.19
peak = 0.73
-0.05
0.00
0.05
0.10
0.15
Time (s)
b
ΔF/F = 1
Amp = 1
Estimated quanta
6
4
2
0
0.2 s
c
Number of observations
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
0.0
0.5
1.0
1.5
2.0
2.5
Event Amplitude (a.u.)
d
Interpeak distance (normalized)
2
1
0
2
3
4
5
6
7
8
9
Peak Number

Figure 3

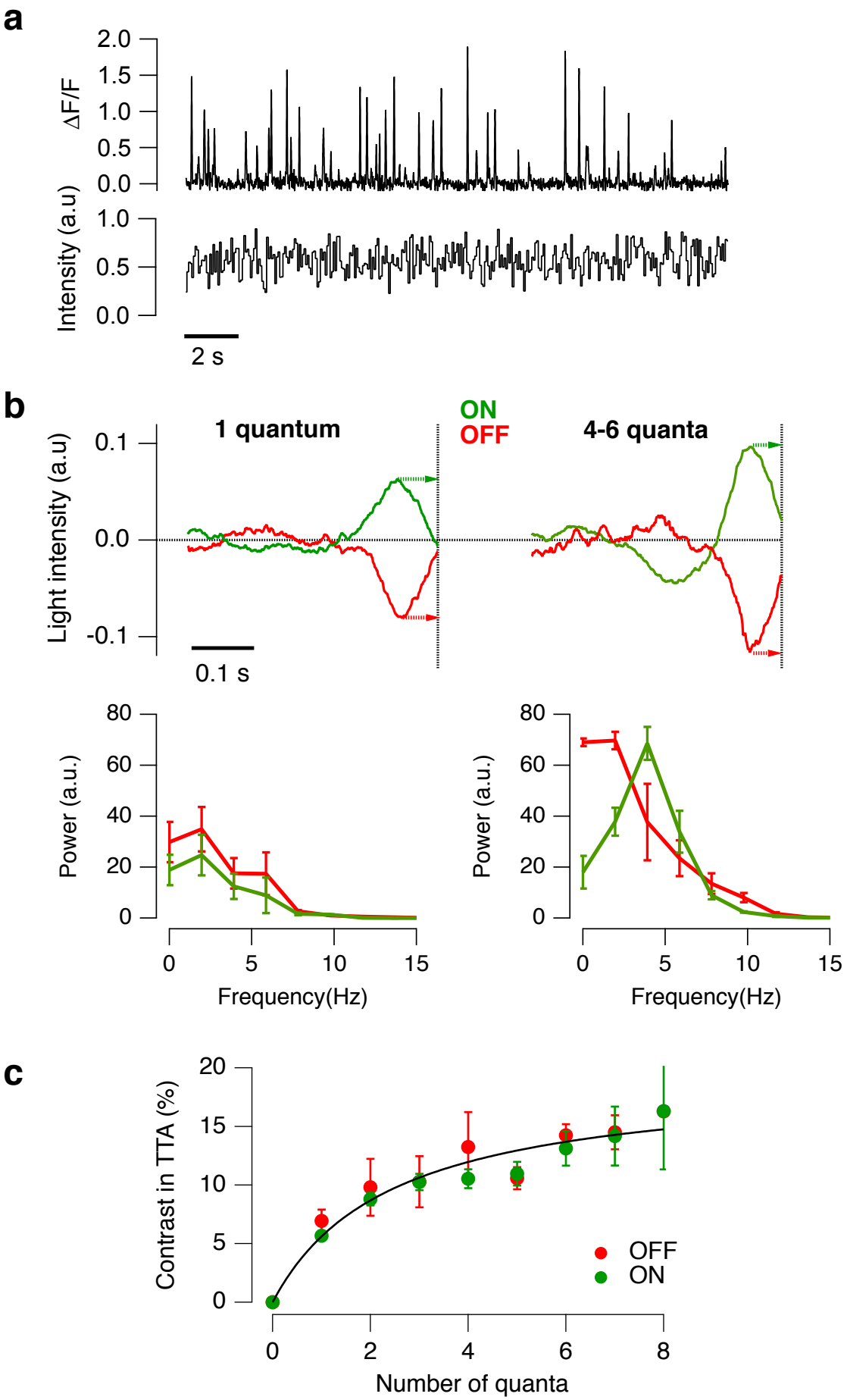
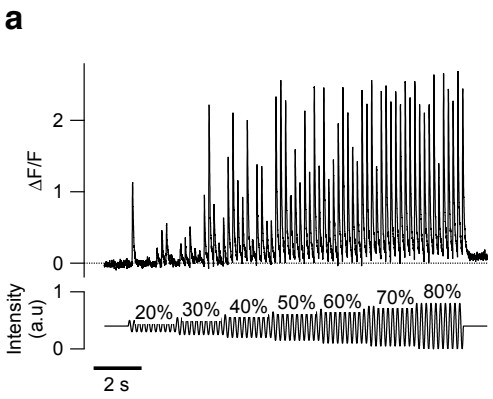
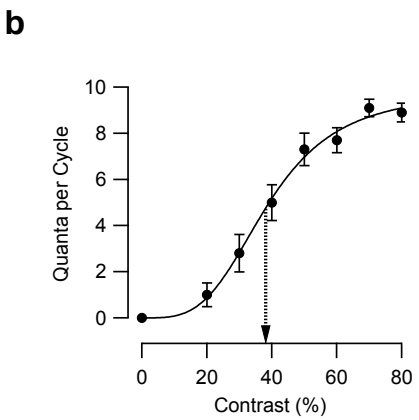
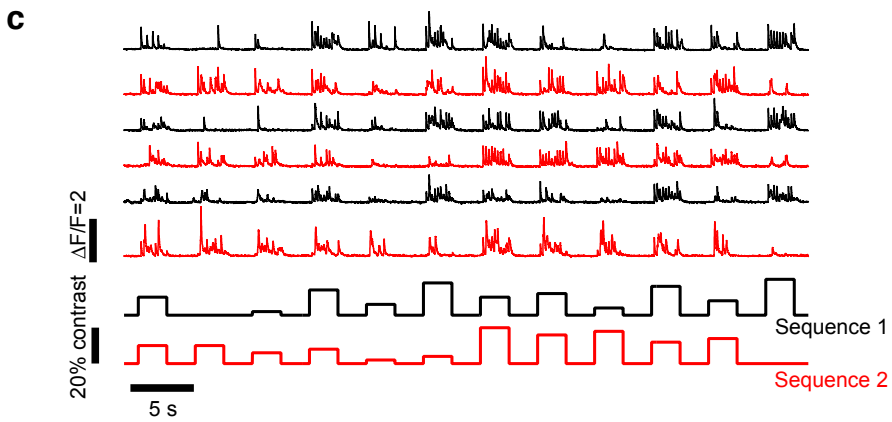
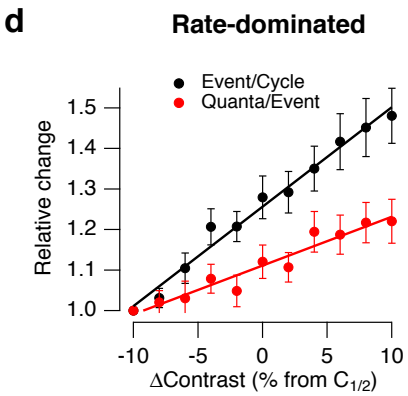

a
2.0
1.5
1.0
0.5
0.0
ΔF/F
1.0
0.5
0.0
Intensity (a.u)
2 s
b
1 quantum
ON
OFF
4-6 quanta
0.1
0.0
-0.1
Light intensity (a.u)
0.1 s
80
60
40
20
0
Power (a.u.)
0
5
10
15
Frequency(Hz)
c
20
15
10
5
0
Contrast in TTA (%)
OFF
ON
0
2
4
6
8
Number of quanta

Figure 4


a
ΔF/F
2
1
0
Intensity (a.u)
1
0
20%
30%
40%
50%
60%
70%
80%
2 s


b
Quanta per Cycle
10
8
6
4
2
0
0
20
40
60
80
Contrast (%)


c
ΔF/F=2
20% contrast
Sequence 1
Sequence 2
5 s


d
Rate-dominated
Event/Cycle
Quanta/Event
Relative change
1.5
1.4
1.3
1.2
1.1
1.0
-10
-5
0
5
10
ΔContrast (% from C1/2)

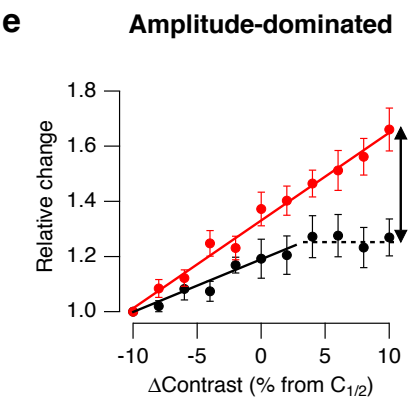
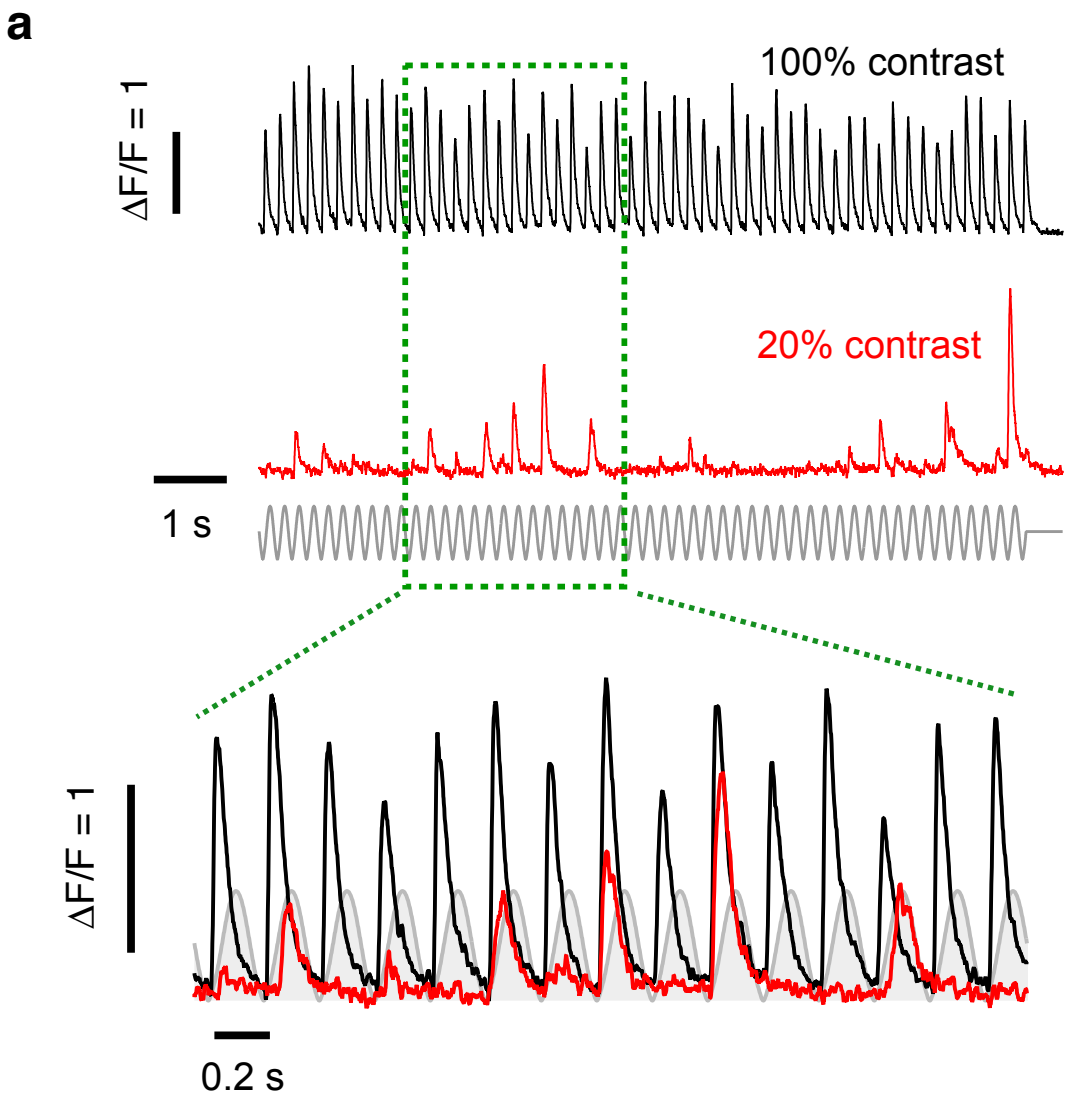

e
Amplitude-dominated
Relative change
1.8
1.6
1.4
1.2
1.0
-10
-5
0
5
10
ΔContrast (% from C1/2)

Figure 5


a
100% contrast
20% contrast
ΔF/F = 1
1 s
ΔF/F = 1
0.2 s

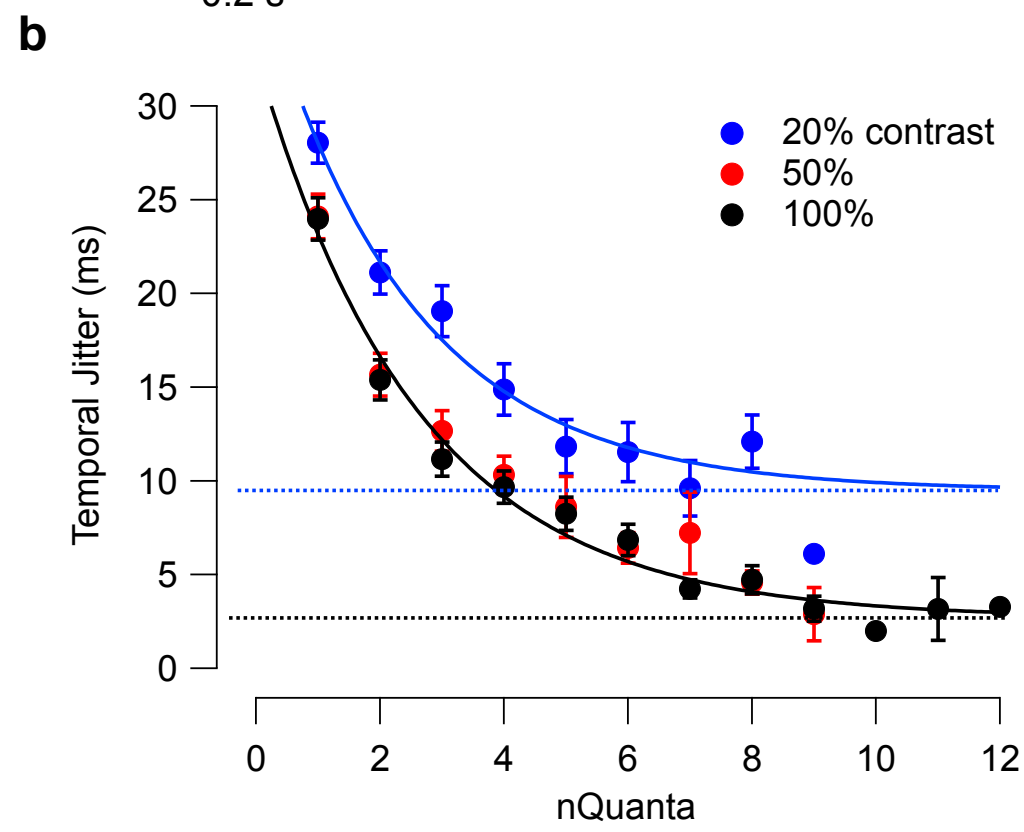
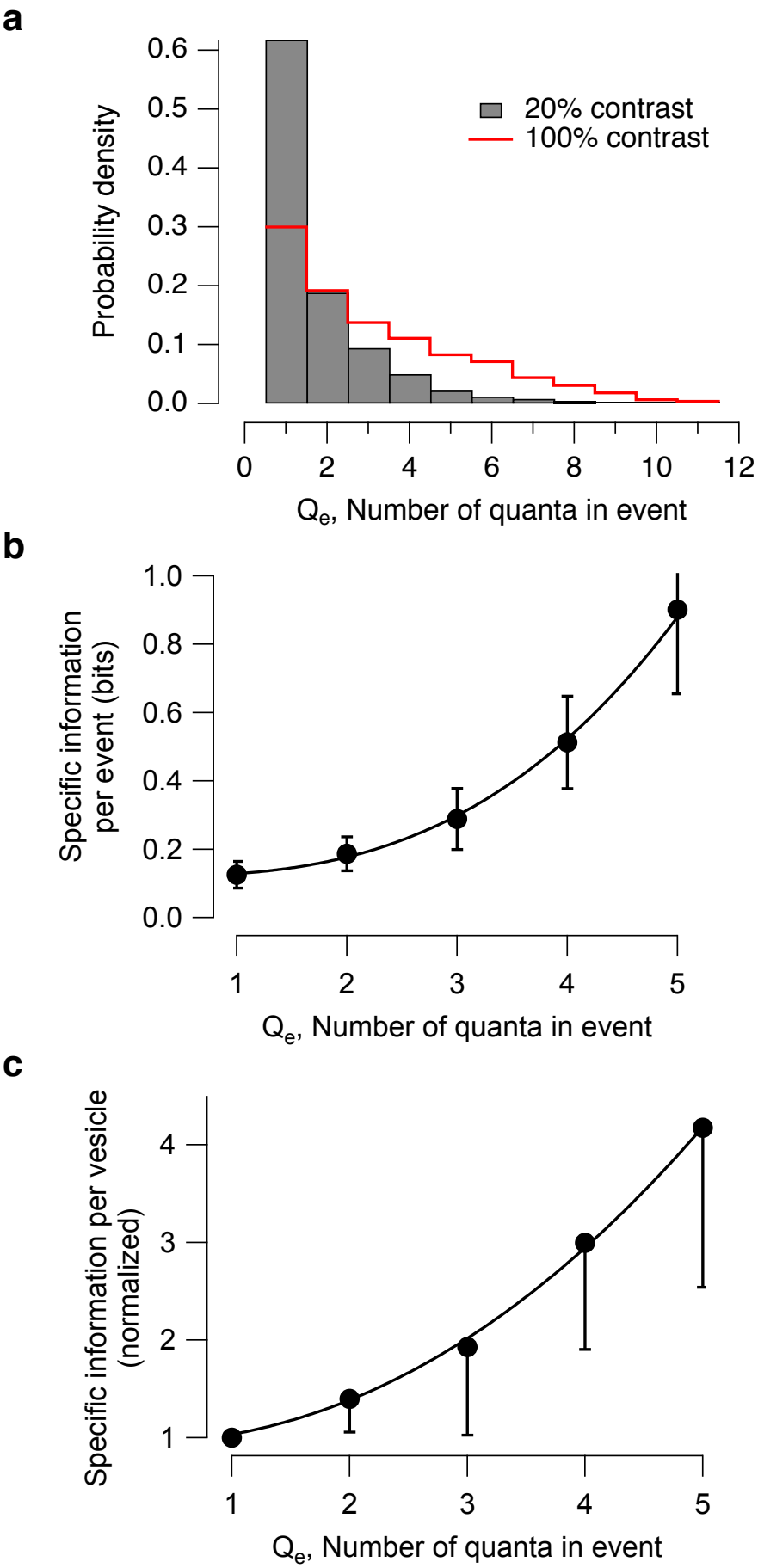

b
20% contrast
50%
100%
Temporal Jitter (ms)
30
25
20
15
10
5
0
0
2
4
6
8
10
12
nQuanta

Figure 6


a
Probability density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20% contrast
100% contrast
0
2
4
6
8
10
12
Qe, Number of quanta in event
b
Specific information per event (bits)
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Qe, Number of quanta in event
c
Specific information per vesicle (normalized)
4
3
2
1
1
2
3
4
5
Qe, Number of quanta in event

Plasmid	Amplicon	5'-Forward Primer-3'	5'-Reverse Primer-3'
Tol2 pDest 10 x UASiGluSnFR (kind gift from Michael Orger Lab)	Whole plasmid	TCCTGTGGTGTCTGAAACACCTGTGCTGCTCGCAGCTGCTGA	GTTAGGGATAACAGGGTAATTCCAAAATCAGCCACAGGATCAAGAGCA
14 x UASiGluSnFR _Mossy Heart	Cmcl2-eGFP-SV40	ATCCTGTGGCTGATTTTGGAATTACCCTGTTATCCCTAACGCC	AGCAGCTGCGAGCAGCACAGGTGTTTCAGACACCACAGGAA
I-SceI pBKS Ribeye SyGCaMP6.10.500 _Bleeding heart	Whole plasmid without SyGCaMP6 cds	GGCGCTCTGGATATGTAGCGGCGGCCGCGACTCTAGATCA	CTTTCAGGAGGCTTGCTTCAGGTGGCTCGAGATCTGAGTCC
Brn3cGal4 (kind gift from Martin Meyer Lab)	Gal4 VP16	GACTCAGATCTCGAGCCACCTGAAGCAAGCCTCCTGAAAGATGAAG	TGATCTAGAGTCGCGGCCGCCGCTACATATCCAGAGCGCC

Supplementary Table 1

Plasmid and primer information

The underlined nucleotide sequences show the part of the primer that anneals to the template during the polymerase chain reaction.